DEUTSCHE
BUNDESBANK
EUROSYSTEM

15th Meeting of the Ottawa Group
10 – 12 May 2017

Session 6: Challenges of “big data”

A “big data” gaze at why electronic transactions and web-scraped data are no panacea

Jens Mehrhoff, Eurostat

Wu (2012) asked: “Is an 80% non-random sample ‘better’ than a 5% random sample in measurable terms? 90%? 95%? 99%?” The answer has been given by Meng (2014); he considers the case of a large but biased administrative record and a small but unbiased random sample from the same population.

His key message is that, as far as statistical inference goes, what makes a “big data” set big is typically not its absolute size, but its relative size to its population. Therefore, the question which data set one should trust more is unanswerable without knowing the population size. But the general message is the same: when dealing with self-reported data sets, do not be fooled by their apparent large sizes or by common wisdom from studying probabilistic samples; or what matters most is the quality, not the quantity.

This paper translates Meng’s approach to electronic transactions and web-scraped data and shows that national statistical institutes should not ignore seemingly tiny probabilistic datasets (e.g. from traditional price collection) when producing consumer price indices.